Medical Knowledge Adaptation for Large Language Models: Strategies and Comparative Analysis

Anonymous ACL submission

Abstract

The development of Large Language Models (LLMs) has led to increased focus on their adaptation to specialized domains and languages, particularly in settings with limited domain-specific data. While recent studies have questioned the benefits of domain-adaptive pre-training (DAPT) in English medical contexts, our work demonstrates that domain adaptation can be effective when strategically implemented. Using French medical domain adaptation as a case study, we systematically evaluate different adaptation strategies: continual pre-training (CPT), supervised fine-tuning (SFT), and combined approaches (CPT followed by SFT). Our study highlights that adapting a general-purpose model with novel domain data leads to significant gains (87% win rate), whereas further adapting models already exposed to similar knowledge offers limited benefits. Moreover, while CPT+SFT achieves the best overall performance, direct SFT emerges as a strong, more computationally efficient alternative.

1 Introduction

Recent advances in Large Language Models (LLMs) have intensified the debate on their adaptation to specialized domains and languages. While various adaptation strategies have been proposed, determining the optimal approach becomes particularly challenging when targeting a specific domain in a language with limited resources. This challenge is exemplified in healthcare and medicine, where most adaptation efforts have focused on English-language models (e.g., BioMistral (Labrak et al., 2024), OpenBioLLM (Pal and Sankarasubbu, 2024).

The development of domain-specialized models in non-English languages faces several unique challenges. First, there is often a scarcity of domain-specific training data. Second, the computational cost of different adaptation strategies must be weighed against their effectiveness. Third, the reliability of evaluation methods, particularly when using translated benchmarks, needs to be assessed, as translation may introduce cultural biases or lose domain-specific nuances. These challenges are particularly evident in specialized domains like medicine, where the accuracy and reliability of the model outputs are crucial, yet domain-specific resources in non-English languages remain limited.

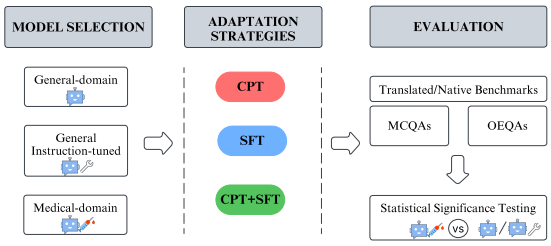


Figure 1: Pipeline for evaluating domain adaptation strategies, showing the three main components: model selection, adaptation strategies (CPT, SFT, and CPT followed by SFT), and evaluation methodology.

Recent work by Jeong et al. (2024) has fundamentally challenged established assumptions regarding domain-adaptive pre-training (DAPT) in medical contexts. Their analysis revealed that medical LLMs adapted through biomedical corpus pre-training frequently fail to demonstrate consistent improvements over their base generic models. These findings raise important questions about the optimal strategy for domain adaptation, particularly when working across languages where the base model's domain knowledge may be less accessible.

To address these challenges, we investigate adaptation strategies for biomedical LLMs, focusing on French—a language that remains underexplored in this domain—while accounting for the limited availability of domain-specific data. We evaluate three distinct approaches: continual pre-training

(CPT), supervised fine-tuning (SFT), and hybrid approaches combining CPT with SFT. We utilize the general-domain Mistral-7B (Jiang et al., 2023) model as our foundation, experimenting with both the base model and its adapted variant on English medical texts, to assess the impact of different initialization points on adaptation effectiveness.

Our evaluation framework encompasses both translated and native French medical datasets. Standard medical benchmarks such as PubMedQA (Jin et al., 2019) and MedMCQA (Pal et al., 2022) are translated for evaluation, while native French medical questions are used to validate the model's capabilities directly. Through both multiple-choice and open-ended question answering tasks, this evaluation approach not only allows us to assess the model's performance on native content but also reveals important methodological insights.

In this paper, we investigate the effectiveness of different strategies for adapting language models to a specific domain in a new language with limited resources. Our contributions can be summarized as follows:

1. We release resources to advance biomedical NLP research in French including medical-adapted models and datasets, all publicly available on HuggingFace[1] under the Apache 2.0 license to support further research.

2. We define an evaluation framework to assess the effectiveness of different adaptation strategies, ensuring comparability across models and approaches.

3. We analyze the impact of various adaptation techniques and provide practical guidelines for selecting the most suitable strategy based on the available training data—whether raw, unannotated text or curated, labeled datasets—and the available computational resources.

2 Related Work

LLM adaptation to the medical domain has seen significant development, driven by the potential to enhance healthcare applications. In this domain, two primary adaptation strategies have emerged. Continual pre-training (CPT) extends the model's pre-training on domain-specific corpora, enabling it to learn domain knowledge while maintaining its general language capabilities. Supervised fine-tuning (SFT), on the other hand, adapts the model through instruction-output pairs, focusing on specific tasks and response formats. CPT has been widely adopted, with models like MediTron (Chen et al., 2023b), BioMistral (Labrak et al., 2024), and PMC-LLaMA (Wu et al., 2023) demonstrating success through adaptation on medical corpora. However, recent work by Jeong et al. (2024) challenges these findings through a more rigorous evaluation methodology: using direct model-to-base comparisons, model-specific prompt optimization, and statistical significance testing. Their methodology revealed that previously reported improvements from medical adaptation were often not statistically significant. Alternative approaches using SFT, as demonstrated by ChatDoctor (Li et al., 2023) and MedAlpaca (Han et al., 2023), have shown promising results in medical tasks through instruction tuning, though these studies also focus on English only.

The challenge of domain adaptation becomes more complex when considering non-English languages, where domain-specific resources are often limited. Recent efforts have addressed this English-centric nature through multilingual approaches. Medical mT5 (García-Ferrero et al., 2024) introduces a text-to-text multilingual model trained on a large corpus spanning English, French, Italian, and Spanish. BiMediX (Pieri et al., 2024) presents a bilingual medical mixture of experts model for English and Arabic, while Apollo (Wang et al., 2024) develops medical LLMs across six languages through the ApolloCorpora dataset and XMed-Bench benchmark. MMedLM (Qiu et al., 2024) provides additional frameworks for multilingual medical adaptation. However, these works primarily rely on translated benchmarks for evaluation, with limited assessment on native language medical tasks raising questions about the models' true capabilities in each target language.

Evaluating medical LLMs presents unique challenges, particularly in multilingual contexts. While benchmarks like PubMedQA (Jin et al., 2019), MedQA (Jin et al., 2019) and MedMCQA (Pal et al., 2022) are widely used, they predominantly serve English-language models. For other languages' evaluation, researchers typically rely on translated benchmarks, with few native language resources. The prevalence of MCQ tasks in these benchmarks also raises questions about comprehen-

[1] https://huggingface.co/Anony-mous123 (will be deanonymized after review)

sive capability assessment.

While these works demonstrate various approaches to medical domain adaptation, there has not been a controlled evaluation of adaptation strategies using a common framework. Previous studies either focus on a single adaptation method or compare models with different architectures and training data, making it difficult to assess the relative effectiveness of CPT versus SFT approaches. Additionally, the trade-offs between computational costs and performance gains remain unclear, particularly in resource-constrained settings. In this work, we address these gaps by conducting a controlled comparison of adaptation strategies using the same base model architecture and evaluation framework, aiming to provide clear guidance for developing medical LLMs in non-English languages.

3 Experimental Framework

We define a framework for evaluating domain adaptation strategies in low-resource settings, as illustrated in Figure 1. Starting from different base models, we investigate various adaptation paths to understand how the choice of starting point and adaptation strategy affects performance and computational efficiency.

Our investigation addresses two key research questions:

RQ1: Does the choice of a base model (general-purpose *vs.* already domain-adapted in English) significantly impact adaptation success?

RQ2: Which adaptation strategy provides the best balance between performance and computational requirements?

To address these questions, we present our experimental methodology in the following sections: base models and adaptation strategies in Section 3.1, training data in Section 3.2, training procedures in Section 3.3), and evaluation protocol in Section 3.4.

3.1 Base Models and Adaptation Approaches

We evaluate adaptation strategies for French biomedical language models using the Mistral-7B architecture family. Our investigation uses three base models, each representing a different starting point for medical domain adaptation:

- **Mistral-7B-v0.1** (Jiang et al., 2023): A 7-billion-parameter general domain LLM.
- **Mistral-7B-instruct-v0.1** (Jiang et al., 2023): The instruction-tuned variant of Mistral-7B-v0.1.
- **BioMistral-7B** (Labrak et al., 2024): An English-based medical domain-adapted variant from Mistral-7B-instruct-v0.1 further pre-trained on PubMed Central Open Access textual data.

The selection of Mistral as our foundation model was motivated by its reasonable French language capabilities compared to other open-source LLMs and its use in comparable studies for English (Labrak et al., 2024). We investigate three distinct adaptation strategies:

- **Continual Pre-training (CPT)**. Further training on domain-specific corpora.
- **Supervised Fine-tuning (SFT)**. Adaptation using instruction-response pairs.
- **CPT+SFT**. A sequential application of CPT followed by SFT.

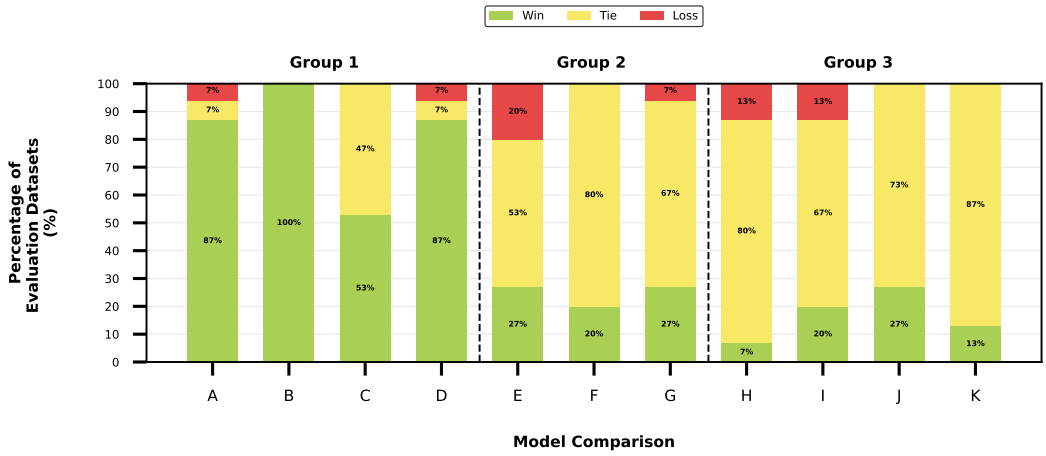
These strategies are applied across different model paths, grouped into families based on their base models, as illustrated in Table 2b.

3.2 Training Data

Our adaptation strategies utilize two distinct datasets.

CPT strategy We employ the NACHOS (openN crAwled frenCh Healthcare cOrpuS) corpus, an open-source French medical dataset spanning 7.4 GB with over one billion words collected from 24 high-quality French-language medical websites (Labrak et al., 2023). Detailed information about the corpus compilation and characteristics is provided in Appendix A.

SFT strategy We constructed a dataset of 30K medical question-answer pairs, equally distributed across three categories: (1) 10K native French medical QAs sourced from medical examinations, including pharmacy specialization and other medical board exams, (2) 10K translated QAs from English medical datasets, covering multiple-choice questions from U.S. medical board exams and medical flashcards, and (3) 10K generated QAs derived from French medical texts, created using a language model and filtered through a multi-step quality assessment process. The dataset includes both multiple-choice questions (MCQs) with single and



(a)

Group	ID	Medical Model	Base Model	Strategy
Group 1 (Mistral)	A	Mistral-7B-Nachos	Mistral-7B-v0.1	CPT
	B	Mistral-7B-Nachos-instruct	Mistral-7B-v0.1	CPT+SFT
	C	Mistral-7B-Nachos-instruct	Mistral-7B-Nachos	CPT+SFT
	D	MedMistral-7B-chat	Mistral-7B-v0.1	SFT
Group 2 (Mistral-Instruct)	E	Mistral-7B-Instruct-Nachos	Mistral-7B-Instruct-v0.1	CPT
	F	Mistral-7B-Instruct-Nachos-instruct	Mistral-7B-Instruct-Nachos	CPT+SFT
	G	Mistral-7B-Instruct-Nachos-instruct	Mistral-7B-Instruct-v0.1	CPT+SFT
Group 3 (BioMistral)	H	BioMistral-Nachos-7B	BioMistral-7B	CPT
	I	BioMistral-Nachos-7B-instruct	BioMistral-7B	CPT+SFT
	J	BioMistral-Nachos-7B-instruct	BioMistral-Nachos-7B	CPT+SFT
	K	BioMistral-7B-chat	BioMistral-7B	SFT

(b)

Figure 2: Evaluation of model adaptation strategies. (a) Win/Tie/Loss analysis across QA datasets, showing the proportion of datasets where each adapted medical model exhibits significant improvement (Win), no significant difference (Tie), or significant degradation (Loss) compared to its base model. Model comparisons are labeled A–K, as shown in (b). (b) Medical models, their base models, and the adaptation strategies used. For CPT+SFT adaptations, we present both comparisons: with the direct base model (after CPT) and with the original base model (before CPT) to evaluate the contribution of each adaptation step.

multiple correct answers, as well as open-ended questions (OEQs) with and without context. Further details on data composition and sources are provided in Appendix C.

3.3 Training Process

Our training procedures employ contrasting adaptation approaches to explore the trade-off between computational efficiency and model plasticity. To examine adaptation strategies at opposite ends of the parameter efficiency spectrum, we apply full fine-tuning for CPT and a parameter-efficient approach for SFT.

CPT strategy We use an improved batching method following BioMistral (Labrak et al., 2024), which utilizes a post-tokenization grouping strategy to aggregate variable-sized sequences marked by end-of-sequence tokens (</s>). This approach effectively fills 2,048-token sequences without the need for padding. The training was conducted for 2.8 epochs with the following setup: the AdamW optimizer (Loshchilov and Hutter, 2019) was used, with a learning rate of 2×10^{-5} and a cosine scheduler, without warmup. The weight decay was set to 0.01, and the batch size was 16 with gradient accumulation steps of 2. The trainings were performed on 32 NVIDIA GPUs either A100 80GB or H100 80GB.

SFT strategy We implement DoRA (Weight-Decomposed Low-Rank Adaptation) (Liu et al., 2024), an enhancement of LoRA (Hu et al., 2021) that decomposes pre-trained weights into magnitude and direction components. This approach aims to achieve fine-tuning capacity while minimizing trainable parameters through LoRA's directional updates. We selected DoRA after conducting preliminary experiments comparing its performance against LoRA and VeRA (Kopiczko et al., 2024),

where DoRA demonstrated superior adaptation efficiency and task-specific performance. SFT training was run for 10 epochs; complete hyperparameter details are provided in Appendix B.

This design choice—full fine-tuning for CPT and parameter-efficient fine-tuning for SFT—allows us to evaluate adaptation strategies that lie at opposite ends of the trade-off between computational cost and model flexibility.

3.4 Evaluation Protocol

This section outlines our evaluation protocol. We introduce a French medical reasoning and knowledge benchmark (native and translated MCQs and OEQs), describe our prompting strategy, and detail the evaluation metrics and significance analysis.

Translated MCQ Datasets The translated datasets consist of French versions of established English medical benchmarks, translated using GPT-3.5-turbo: MedQA (Jin et al., 2020), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), and MMLU (Hendrycks et al., 2021) medical subcategories, as detailed in Table 1.

Native French MCQ Datasets For native French evaluation, we use FrBMedQA dataset (Kaddari and Toumi, 2022), which contains questions from French biomedical Wikipedia articles across eight Unified Medical Language System (UMLS) semantic groups: chemicals and drugs, anatomy, physiology, disorders, phenomena, procedures, genes and molecular sequences, and devices. The questions were converted from close-style to multiple-choice format using GPT-4o-mini (prompt details in Appendix D). We additionally scraped and processed FrMedMCQA from S-Editions[2], a platform offering medical resources and study materials for medical students in France. The data collection involved automated extraction followed by manual cleaning to ensure question-answer pair quality. The final dataset covers oncology, cardiovascular medicine, dermatology, endocrinology, gynecology, hematology, infectious diseases, neurology, ophthalmology, pediatrics, psychiatry, and rheumatology.

OEQ Datasets For OEQ evaluation, we translated the K-QA dataset (Manes et al., 2024) to French using GPT-4o-mini. This dataset contains 201 patient questions from K Health[3] platform conversations, answered by in-house physicians. We also scraped and processed two native datasets from S-Editions: FrClinicalQA with clinical case questions in cardiology, oncology, pneumology, infectious diseases, endocrinology, and rheumatology; and FrMedQA with medical questions without clinical context in these domains. Each scraped dataset underwent automated cleaning and manual verification.

An overview of dataset sources and sizes is listed in Table 1.

QA type	Dataset	Context	Test set
Translated MCQ	MedQA (4 & 5 Options)	✗	1,273
	MedMCQA	✗	4,183
	PubMedQA	✓	500
	MMLU: Anatomy	✗	135
	MMLU: Clinical Knowledge	✗	265
	MMLU: College Biology	✗	144
	MMLU: College Medicine	✗	173
	MMLU: Professional Medicine	✗	272
	MMLU: Medical Genetics	✗	100
Native MCQ	FrBMedQA	✗	2,156
	FrMedMCQA	✗	183
Translated OEQ	K-QA	✗	201
Native OEQ	FrClinicalQA	✓	262
	FrMedQA	✗	81

Table 1: Table 1: Evaluation datasets categorized by QA type, source language(translated or native French), context availability, and test set size.

Prompting Strategy We conduct zero-shot evaluation to simulate real-world scenarios and due to dataset constraints, as most datasets except translated MCQs consist only of small test sets (Table 1). To generate responses, we employ a greedy decoding strategy. For MCQ tasks, following Liang et al. (2022), Beeching et al. (2023) and Chen et al. (2023a), we filter the vocabulary to include only tokens (choice letters) corresponding to the expected answer options, preventing the model from generating irrelevant tokens or hallucinations.

As a supplementary experiment, we investigate whether English examples, matching the models' initial training language, affect performance on French medical questions. We implement 3-shot in-context learning on the translated MCQ datasets, using three sets of randomly selected examples from each dataset's training set. We test two configurations: (1) French prompts and questions with English in-context examples, where the model is exposed to French-language tasks but provided with English-language examples, and (2) all-French prompts, questions, and examples, ensuring that all input is in French. Complete prompting templates for both zero-shot and few-shot evaluations are detailed in Appendix E.

[2] https://s-editions.fr/
[3] https://khealth.com/

Evaluation Metrics For MCQ tasks, we report the exact-match accuracy and for OEQ tasks, we evaluate performance using the F1 BERTScore (Zhang et al., 2020).

Statistical Significance Assessment To determine whether the observed performance improvements from adaptation are statistically significant, we employ a percentile bootstrap method, similar to Jeong et al. (2024). This approach involves resampling (with replacement) from the test set to create samples of the same size as the original. For each resample, we compute the performance difference (e.g., accuracy for MCQ or F1 BERTScore for OEQ) between paired models. This process is repeated 10,000 times, generating a distribution of relative performance metrics. A 95% confidence interval is then derived from this distribution, and a difference is considered statistically significant if the interval does not include zero.

Unlike Jeong et al. (2024), we applied the Bonferroni correction to account for the multiple comparisons conducted in our study. This correction mitigates the increased likelihood of Type I errors (false positives) that arise when testing multiple hypotheses. The Bonferroni correction adjusts the significance threshold to α/m, where α is the desired overall significance level (0.05 in our case) and m is the total number of comparisons.

To evaluate the generality of the adapted models, we analyze performance differences at the dataset level, enabling us to compute the win rate, that is the proportion of datasets where a given model outperforms its base version.

4 Results and Discussion

In this section, we present our analysis of the adaptation strategies described in Section 3, providing answers to our research questions RQ1 (impact of base model selection), RQ2 (effectiveness of adaptation strategies), and RQ3 (reliability of evaluation methodologies). Results from our few-shot evaluation experiments are presented in Appendix F, as they provide supplementary insights but do not affect our main findings about adaptation strategies.

4.1 Base Model Selection

We present our evaluation results across three model groups on multiple tasks. Tables 2, 3, and 6 show performance on translated MCQs, native MCQs, and OEQs respectively.

Model Performance Analysis

Group 1 (Mistral-based) Starting from a general-purpose model (Mistral-7B-v0.1), Mistral-7B-Nachos-instruct demonstrates substantial improvements across all metrics. On translated MCQs, performance increases from 0.87% to 47.83%, while native MCQs show improvement from 3.97% to 36.55%. For OEQs, F1 BERTScore improves from 0.55 to 0.67. These gains are statistically significant with a 100% win rate over the baseline.

Group 3 (BioMistral-based) Starting from an English medical model shows contrasting results. On translated MCQs, we observe performance degradation across all adaptations. Native MCQs show improvement from 30.13% to 35.52% (BioMistral-Nachos-7B-instruct), though not statistically significant. The most notable result appears in OEQs, where adaptation achieves a statistically significant improvement of 0.22 in F1 BERTScore.

Group 2 (Mistral-instruct-based) The instruction-tuned starting point yields intermediate results. Mistral-7B-Instruct-Nachos-instruct improves from 39.96% to 43.03% on translated MCQs, from 27.13% to 36.46% on native MCQs, and maintains a 0.67 F1 BERTScore on OEQs. Despite these apparent improvements, statistical testing reveals low significance with only a 27% win rate.

Impact of Base Model Selection Statistical comparison between the best performing models from each group (Table 5) shows Mistral-7B-Nachos-instruct (Group 3) significantly outperforming BioMistral-Nachos-7B-instruct (Group 1) with a 73% win rate. This suggests that starting from a general-purpose model proves more effective than building upon an already medical-specialized model. The comparison with Mistral-7B-Instruct-Nachos-instruct (Group 2) shows mixed results (40% wins, 60% ties), indicating that while starting from a general model might be advantageous over an instruction-tuned variant, the benefits are less pronounced. These results can be explained by several factors. The limited gains in Group 1 (BioMistral-based) suggest that when a model has already acquired medical knowledge during English pre-training, further adaptation on French medical data may be redundant or even detrimental for factual knowledge tasks. This is evidenced by BioMistral's strong base performance but limited gains from adaptation. The intermediate perfor-

Model	Strategy	Average	PubMedQA	MedQA 4 Options	MedQA 5 Options	MedMCQA	MMLU					
							Clinical Knowledge	Medical Genetics	Anatomy	Pro. Medicine	College Biology	College Medicine
Mistral-7B-v0.1	Base Model	0.87	4.00	0.08	0.24	0.19	1.51	0.00	0.00	0.00	2.08	0.58
Mistral-7B-Nachos	CPT	36.00	41.00	28.83	23.10	33.09	44.15	43.00	37.78	27.21	46.53	35.26
Mistral-7B-Nachos-instruct	CPT+SFT	**47.83**	**64.00**	**42.66**	**35.19**	**37.99**	**49.43**	**56.00**	42.96	**53.68**	**47.22**	49.13
MedMistral-7B-chat	SFT	43.65	54.60	39.67	31.89	35.79	44.91	43.00	**45.19**	48.53	40.28	**52.60**
Mistral-7B-Instruct-v0.1	Base Model	39.96	54.40	29.14	24.90	31.87	46.42	44.00	37.78	**46.32**	40.28	44.51
Mistral-7B-Instruct-Nachos	CPT	43.03	34.80	**37.08**	**32.29**	**38.42**	**50.57**	**59.00**	**42.96**	40.81	**49.31**	**45.09**
Mistral-7B-Instruct-Nachos-instruct	CPT+SFT	**43.20**	**59.20**	36.29	31.50	36.39	42.26	53.00	42.22	40.07	46.53	44.51
BioMistral-7B	Base Model	**41.39**	**54.60**	32.44	26.08	31.68	**52.08**	43.00	40.74	**45.22**	**42.36**	**45.66**
BioMistral-Nachos-7B	CPT	35.14	14.80	28.75	27.49	30.62	44.15	42.00	39.26	43.38	41.67	39.31
BioMistral-Nachos-7B-instruct	CPT+SFT	34.53	36.60	35.59	**30.24**	**35.41**	32.83	37.00	34.07	29.41	38.89	35.26
BioMistral-7B-chat	SFT	37.68	44.60	**36.68**	**30.24**	31.87	38.49	**44.00**	**41.48**	39.34	35.42	34.68

Table 2: Zero-shot performance on translated MCQ tasks. Scores are reported using exact-match accuracy. The best-performing model within each group is highlighted in **bold**, and the overall best-performing model is underlined.

Model	Strategy	Average	FrBMedQA	FrMedMCQA
Mistral-7B-v0.1	Base Model	3.97	7.93	0.00
Mistral-7B-Nachos	CPT	33.48	50.56	16.39
Mistral-7B-Nachos-instruct	CPT+SFT	**36.55**	**50.70**	**22.40**
MedMistral-7B-chat	SFT	29.88	48.28	11.47
Mistral-7B-Instruct-v0.1	Base Model	27.13	43.88	10.38
Mistral-7B-Instruct-Nachos	CPT	**36.46**	**53.25**	19.67
Mistral-7B-Instruct-Nachos-instruct	CPT+SFT	35.50	50.79	**20.21**
BioMistral-7B	Base Model	30.13	46.06	14.20
BioMistral-Nachos-7B	CPT	32.83	**47.63**	18.03
BioMistral-Nachos-7B-instruct	CPT+SFT	**35.52**	47.54	**23.49**
BioMistral-7B-chat	SFT	27.93	46.57	9.28

Table 3: Zero-shot performance on native French MCQA tasks. Scores represent exact-match accuracy. The best model in each group is highlighted in **bold** and the best model overall is underlined.

mance of Group 2 suggests that while instruction tuning provides some benefits, it may constrain the model's ability to fully adapt to new domains compared to starting from a general model.

4.2 Adaptation Strategy Effectiveness

Having established Group 1 (Mistral-based models) as the most effective starting point, we focus our analysis of adaptation strategies within this group where improvements are statistically significant and meaningful.

CPT CPT alone (Mistral-7B-Nachos) shows significant improvements: 11.83% and 29.51% increases on translated and native MCQs respectively, while maintaining baseline performance on OEQs.

CPT+SFT The addition of SFT enhances these gains, with Mistral-7B-Nachos-instruct achieving 47.83% on translated MCQs, 36.55% on native MCQs, and 0.67 F1 BERTScore on OEQs.

SFT Direct SFT (MedMistral-7B-chat) demonstrates strong results with 43.65% on translated MCQs, 29.88% on native MCQs, though showing slight degradation on OEQs from 0.55 to 0.52.

Strategies	Win	Tie	Loss
CPT+SFT vs. CPT	**0.67**	0.33	0
CPT+SFT vs. SFT	**0.4**	0.54	0.06

Table 4: Statistical significance comparison of adaptation strategies. Win/Tie/Loss rates indicate the proportion of datasets where CPT+SFT shows significant improvement/no significant difference/significant degradation compared to CPT-only and SFT-only adaptations in Group 1 (Mistral-based models)

Groups	Win	Tie	Loss
Group 1 vs. Group 3	**0.73**	0.27	0
Group 1 vs. Group 2	**0.4**	0.6	0

Table 5: Statistical significance comparison between groups to to determine the optimal starting point for adaptation. Win/Tie/Loss rates compare the best performing models from Group 1 against those from Group 2 and Group 3.

Strategies Comparison Statistical testing between strategies (Table 4) confirms CPT+SFT's advantages over CPT (67% wins, 33% ties) but shows less consistent superiority over SFT (40% wins, 54% ties, 6% losses). While CPT+SFT achieves the best overall performance, direct SFT offers a compelling alternative when considering computational efficiency, as we discuss in the following section (Section 4.3).

4.3 Computational Efficiency and Environmental Impact

Our analysis of adaptation strategies considers not only performance but also computational costs and environmental impact. We assess these factors using three metrics: training time, carbon emissions (kgCO2e), and monetary costs.

The CPT approach, while effective, demands substantial resources. Based on internal estima-

tions, training on 7.4GB of medical data requires 32 GPUs (NVIDIA H100 or A100), generating approximately 9–10 kgCO2e per adaptation. In contrast, SFT processes only 36MB of data and runs on 1–2 GPUs, reducing emissions to 2.5–2.6 kgCO2e. The combined CPT+SFT approach accumulates the costs of both stages, resulting in total emissions of 10–12 kgCO2e.

Monetary costs follow a similar trend: CPT training ranges from 590 USD[4] to 1,073 USD depending on the GPU type, while direct SFT costs only 43–45 USD. These findings highlight direct SFT as a significantly more resource-efficient adaptation path, requiring just 25% of the computational resources and carbon emissions of CPT or combined approaches. Further details on training time, carbon emissions, and monetary costs are provided in Appendix H.

Our findings present an interesting parallel to recent work Jeong et al. (2024) questioning the effectiveness of medical domain adaptation. While they found that general English models already possess strong medical capabilities, making additional medical training redundant due to exposure to medical data (PubMed) during pre-training, our results with BioMistral (Group 3) show similarly limited gains from additional medical adaptation. However, our Group 1 results reveal that when starting from a general model and adapting with new medical data in the target language (here French), domain adaptation can provide significant benefits (87% win rate over baseline). This suggests that the effectiveness of domain adaptation may depend on both the starting point and whether the base model has already been exposed to similar domain-specific data during pre-training, directly addressing RQ1 about the impact of base model selection on adaptation success.

Furthermore, while CPT+SFT achieves the best performance in this setting, our analysis shows that direct SFT offers a compelling alternative when computational resources are limited. With just 25% of the computational cost and carbon emissions of CPT or CPT+SFT, direct SFT delivers substantial improvements at a fraction of the resource requirements. This highlights an important trade-off between performance gains and efficiency, suggesting that in resource-constrained scenarios, direct SFT can be a practical and effective adaptation strategy, thereby answering RQ2 regarding adaptation strategy effectiveness.

Additionally, our evaluation methodology highlights important considerations about assessing domain-adapted models. The divergent performance patterns between multiple-choice and open-ended tasks (in group 3) raise important questions about evaluation methodology, suggesting that current metrics, such as BERTScore, may not effectively distinguish between improvements in factual knowledge versus language generation capabilities.

Model	Strategy	Average	FrClinicalQA	FrMedQA	K-QA
Mistral-7B-v0.1	Base Model	0.55	0.35	0.61	0.70
Mistral-7B-Nachos	CPT	0.55	0.56	0.59	0.51
Mistral-7B-Nachos-instruct	CPT+SFT	**0.67**	**0.65**	**0.64**	0.72
MedMistral-7B-chat	SFT	0.52	0.20	0.62	**0.73**
Mistral-7B-Instruct-v0.1	Base Model	**0.67**	**0.65**	**0.67**	0.70
Mistral-7B-Instruct-Nachos	CPT	0.60	0.51	0.65	0.63
Mistral-7B-Instruct-Nachos-instruct	CPT+SFT	**0.67**	**0.65**	0.64	**0.71**
BioMistral-7B	Base Model	0.44	0.18	0.52	0.63
BioMistral-Nachos-7B	CPT	0.47	0.45	0.60	0.35
BioMistral-Nachos-7B-instruct	CPT+SFT	**0.66**	**0.64**	**0.63**	**0.71**
BioMistral-7B-chat	SFT	0.57	0.39	**0.63**	0.70

Table 6: Zero-shot performance on OEQ tasks. Scores represent f1 BERTScore. The best model in each group is highlighted in **bold** and the best model overall is underlined.

5 Conclusion

This work investigates adaptation strategies for the development of models in the French medical language, providing information on domain adaptation in low-resource settings. Our results tend to show that the effectiveness of domain adaptation depends on the base model's prior exposure to domain knowledge. Our analysis of different adaptation strategies reveals that while combined CPT+SFT achieves the best performance across all tasks, direct SFT offers a compelling alternative, achieving strong results with significantly lower computational requirements. This finding has important implications for resource-efficient domain adaptation.

These findings contribute to our understanding of cross-lingual domain adaptation and provide practical guidelines for developing specialized language models in resource-constrained settings. Future work should focus on developing more efficient adaptation strategies and more reliable evaluation methodologies for assessing domain-specific capabilities.

[4]Estimations from our cloud provider.

6 Limitations

Our evaluation of adaptation strategies faces several limitations. First, due to the scarcity of native French medical evaluation datasets, we rely heavily on translated benchmarks. While we include native French tests, a more comprehensive evaluation would require larger native datasets across diverse medical specialties.

Second, our assessment of model performance on OEQs uses BERTScore, which may not fully capture the medical accuracy of generated responses. The development of specialized metrics for evaluating medical language generation, particularly for non-English languages, remains an important challenge.

Third, while we demonstrate the efficiency of SFT compared to CPT in terms of computational resources, our analysis does not account for the human effort required to create high-quality instruction-tuning datasets. This consideration is particularly relevant for low-resource settings where creating domain-specific instruction data may be costly.

Fourth, our experiments are conducted on the Mistral-7B family of models and may lead to different conclusions if run on different-sized models, or models trained on a substantially different data mixture. Also, datasets we base our conclusions on are mostly question-answering oriented, and not representative of the variety of uses of LLMs possible in the medical domain. Our conclusions might be affected by more diverse evaluation tasks.

Finally, our findings about the effectiveness of adaptation strategies are specific to the medical domain and French language. The generalizability of these results to other domains or languages, particularly those with different resource constraints or linguistic characteristics, requires further investigation.

References

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard hugging face. *Récupérée mai*, 24:2024.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023a. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023b. Meditron-70b: Scaling medical pretraining for large language models. *Preprint*, arXiv:2311.16079.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Iker García-Ferrero, Rodrigo Agerri, Aitziber Atutxa Salazar, Elena Cabrio, Iker de la Iglesia, Alberto Lavelli, Bernardo Magnini, Benjamin Molinet, Johana Ramirez-Romero, German Rigau, Jose Maria Villa-Gonzalez, Serena Villata, and Andrea Zaninello. 2024. Medical mt5: An open-source multilingual text-to-text llm for the medical domain. *Preprint*, arXiv:2404.07613.

Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K. Bressem. 2023. Medalpaca – an open-source collection of medical conversational ai models and training data. *Preprint*, arXiv:2304.08247.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Daniel P. Jeong, Pranav Mani, Saurabh Garg, Zachary C. Lipton, and Michael Oberst. 2024. The limited impact of medical adaptation of large language and vision-language models. *Preprint*, arXiv:2411.08870.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Preprint*, arXiv:2009.13081.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *Preprint*, arXiv:1909.06146.

Zakaria Kaddari and Bouchentouf Toumi. 2022. Frbmedqa: the first french biomedical question answering dataset. *IAES International Journal of Artificial Intelligence (IJ-AI)*, 11:1588.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. *Preprint*, arXiv:2405.01535.

Dawid J. Kopiczko, Tijmen Blankevoort, and Yuki M. Asano. 2024. Vera: Vector-based random matrix adaptation. *Preprint*, arXiv:2310.11454.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Béatrice Daille, Pierre-Antoine Gourraud, Emmanuel Morin, and Mickael Rouvier. 2022. FrenchMedMCQA: A French Multiple-Choice Question Answering Dataset for Medical domain. In *LOUHI 2022*, Abou Dhabi, United Arab Emirates.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. Drbert: A robust pre-trained model in french for biomedical and clinical domains. *Preprint*, arXiv:2304.00958.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Preprint*, arXiv:2303.14070.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation. *Preprint*, arXiv:2402.09353.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Itay Manes, Naama Ronn, David Cohen, Ran Ilan Ber, Zehavi Horowitz-Kugler, and Gabriel Stanovsky. 2024. K-qa: A real-world medical qa benchmark. *Preprint*, arXiv:2401.14493.

Ankit Pal and Malaikannan Sankarasubbu. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. `https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B`. Hugging Face model repository.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering. *Preprint*, arXiv:2203.14371.

Sara Pieri, Sahal Shaji Mullappilly, Fahad Shahbaz Khan, Rao Muhammad Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. 2024. BiMediX: Bilingual medical mixture of experts LLM. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16984–17002, Miami, Florida, USA. Association for Computational Linguistics.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *Preprint*, arXiv:2402.13963.

Xidong Wang, Nuo Chen, Junyin Chen, Yan Hu, Yidong Wang, Xiangbo Wu, Anningzhe Gao, Xiang Wan, Haizhou Li, and Benyou Wang. 2024. Apollo: An lightweight multilingual medical llm towards democratizing medical ai to 6b people. *Preprint*, arXiv:2403.03640.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Towards building open-source language models for medicine. *Preprint*, arXiv:2304.14454.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

A NACHOS Corpus Description

The NACHOS corpus is a French medical open-source dataset compiled through extensive web crawling and text collection. The corpus spans 7.4 GB of data and contains over one billion words (1,088,867,950 words) sourced from 24 French-speaking high-quality websites (Labrak et al., 2023).

Note: Full details of the corpus compilation and processing are available in the original paper (Labrak et al., 2023).

A.1 Corpus Composition

The NACHOS corpus encompasses a diverse range of medical textual sources, including:

- Descriptions of diseases and conditions
- Treatment and medication information
- General health-related advice
- Official scientific meeting reports
- Anonymized clinical cases
- Scientific literature
- Theses
- French translation pairs
- University health courses

A.2 Data Sources

The corpus integrates data from multiple sources, with the most significant contributions coming from:

- HAL (638,508,261 words)
- Haute Autorité de Santé (HAS) (113,394,539 words)
- Drug leaflets (74,770,229 words)
- Medical Websites Scraping (60,561,495 words)
- ANSES SAISINE (51,372,932 words)
- Public Drug Database (BDPM) (48,302,695 words)

A.3 Corpus Preparation

The researchers employed several preprocessing steps:

1. Text collection through web scraping, raw textual sources, and optical character recognition (OCR)
2. Sentence splitting using heuristic methods
3. Aggressive filtering to remove short or low-quality sentences
4. Language classification using a custom classifier trained on multilingual corpora

B SFT hyperparameters

Parameter	Value
Rank	16
LoRA Aplha	16
LoRA Dropout	0.05
Learning rate	2e-05
Train batch size	4
Evaluation batch size	8
Seed	42
Number of GPU	1 NVIDIA H100 80GB Or 2 NVIDIA L40 48GB
Gradient accumulation steps	2
Optimizer	AdamW
Scheduler	Cosine
Number of epochs	10
Target Modules	QKVOGUD

Table 7: Hyperparameters for the Supervised FineTuning (SFT) training

C SFT Training dataset

The Supervised Fine-tuning dataset comprises 30,000 question-answer pairs sourced from three distinct categories: native French medical content, translated English medical content, and generated questions from French medical texts.

C.1 Native French Content

We randomly sampled 10,000 question-answer pairs from two primary sources. The first source is FrenchMedMCQA (Labrak et al., 2022), a dataset containing 3,105 questions derived from French pharmacy specialization diploma examinations. These questions encompass both single and multiple-answer formats, reflecting real examination conditions and standards.

The second source consists of two complementary datasets hosted on Hugging Face[5] : mlabonne/medical-mqca-fr[6] and mlabonne/medical-cases-fr[7]. These datasets consist of multiple-choice questions and clinical case studies sourced from French medical examination databases, encompassing a wide range of medical specialties, including addictology, gerontology, neurology, and psychiatry, among others.

[5] https://huggingface.co/

[6] https://huggingface.co/datasets/mlabonne/medical-mqca-fr

[7] https://huggingface.co/datasets/mlabonne/medical-cases-fr

C.2 Translated Content

Another 10,000 question-answer pairs were sampled from English medical resources and translated to French using jsontt[8], an open-source command-line interface tool that leverages multiple translation services. The source material included the training set of MedQA (Jin et al., 2020), which comprises multiple-choice questions from U.S. medical board examinations, and the Medical Meadow Medical Flashcards compiled by MedAplaca (Han et al., 2023), which cover fundamental medical subjects including anatomy, physiology, pathology, and pharmacology.

C.3 Generated Content

The final 10,000 pairs were generated using a two-phase process.

Initially, we used Mistral-7b-instruct-v0.2 (Jiang et al., 2023) to generate question-answer pairs from contexts extracted from the French subset of Antidote corpus (García-Ferrero et al., 2024). We instructed the model to create question-answer pairs based on provided medical contexts using this prompt Figure 3. To ensure the output was in JSON format, we used Outlines[9], a Python library that guides the generation process so that the output adheres to a specified JSON schema.

Generation Prompt

Vous êtes médecin et votre tâche consiste à fournir une paire de question-réponse en français à partir du contexte suivant :
Contexte : {{context}}
N'oubliez pas de répondre en français!

Figure 3: Instruction template used for generating question-answer pairs in French, based on the given context.

The quality of generated pairs underwent an evaluation using three large language models: Prometheus-7B-v2.0 (Kim et al., 2024), Meta-Llama-3-70B-Instruct (Dubey et al., 2024), and GPT-4o (Hurst et al., 2024). Each model independently scored the pairs on a five-point scale based on relevance, accuracy, and comprehensiveness following the prompt Figure 4. Only pairs receiving scores of 4 or 5 from all three evaluating models were retained for the training corpus, ensuring high-quality training data.

Evaluation Prompt

You are a medical evaluator tasked with assessing question-answer pairs within a given context. Provide a score from 1 to 5 based on the provided score criteria.
[SCORE]: *(score from 1 to 5)*
Do not include any other opening, closing, or explanations.
Score criteria:

- ***Score 1:*** *The question-answer pair is completely irrelevant or incorrect given the context. The answer has major factual errors.*
- ***Score 2:*** *The question is somewhat relevant but the answer has significant inaccuracies or lacks important details from the context.*
- ***Score 3:*** *The question is relevant and the answer is mostly accurate but contains some minor factual errors or omissions.*
- ***Score 4:*** *The question is clear and relevant, and the answer is accurate based on the context with only very minor omissions.*
- ***Score 5:*** *The question is clear, relevant, and the answer is completely accurate and comprehensive based on the given context.*

Remember, your score should consider both the relevance of the context to the medical domain and the accuracy of the question-answer pair. Here's your question-answer pair given the context:
Context : {{context}}
Question : {{question}}
Answer : {{answer}}
[SCORE]:

Figure 4: Instruction template used for evaluating question-answer pairs based on medical relevance and accuracy.

[8] https://github.com/mololab/json-translator
[9] https://github.com/dottxt-ai/outlines

D System prompt for reformulating FrBMedQA dataset

System Prompt

You are a medical question generation assistant. Given the following passage and a question based on it, transform the question into a valid multiple-choice question (MCQ). The MCQ should:

- *Focus on the placeholder by asking specifically about the information that corresponds to it.*
- *The question should not contain @placeholder*
- *The choices in the MCQ should be taken directly from the 'entities_list' and be formatted as options A, B, C, etc.*
- *The question should be phrased in a formal, clear, and precise manner, as a medical expert would phrase it.*
- *The MCQ should not contain any reference to the passage, such as "according to the passage" or "as stated in the passage". The question should be able to stand alone and should not explicitly refer to the passage.*
- *Provide one correct answer, which should correspond to the letter in the MCQ options.*
- *The MCQ should be written in French.*
- *Return the MCQ in json format*

Figure 5: System prompt given to GPT-4o-mini for generating multiple-choice questions in French from given passages and questions.

E Prompt templates used in evaluation in zero-shot and few-shot settings for MCQ and OEQ tasks

We employed a standardized prompt template (Figure 6) across all multiple-choice question (MCQ) evaluations, with the exception of FrMedMCQA, which uniquely features questions with multiple correct answers and required a specialized prompt (Figure 7).

Instruction Template

Nous vous présentons une question scientifique, (un contexte) et (quatre/cinq) choix de réponse. Votre tâche est de trouver la réponse correcte en vous basant sur des faits scientifiques, vos connaissances et votre raisonnement (le contexte fourni). Générez uniquement l'une des lettres suivantes : A, B, C, D, (E). Chaque question n'a qu'une seule réponse. Les justifications ne sont pas permises.
Voici quelques exemples pour vous aider à mieux comprendre la tâche :
{% for i, shot in fewshots.items() %}
Exemple {{i}}:
Context: *{{context}}*
Question: *{{question}}*
Choix:
{% for letter, option in options.items() %}
{{letter}}: {{option}}
{% endfor %}
Réponse: *{{correct_letter}}*
{% endfor %}
Maintenant, répondez à cette question (en vous basant sur le contexte):
Contexte: *{{context}}*
Question: *{{question}}*
Choix:
{% for letter, option in options.items() %}
{{letter}}: {{option}}
{% endfor %}
Réponse :

Figure 6: Instruction template for zero-shot and few-shot evaluations. For zero-shot evaluations, the few-shot examples are omitted.

The base template was dynamically modified according to specific corpus characteristics:

- If the corpus included a context, the placeholder (un contexte) was replaced with the actual context text. For corpora without context, this part was omitted.
- The number of answer choices varied depending on the corpus, with (quatre/cinq) replaced

by "quatre" (four) or "cinq" (five) as appropriate.

- The letter (E) was included for corpora with five options and omitted otherwise.

- (en vous basant sur le contexte): This placeholder was included for corpora that provided context and omitted otherwise.

Instruction Template

Nous vous présentons une question scientifique suivie de plusieurs choix de réponse. Votre tâche est de sélectionner la ou les lettres correspondant aux réponses correctes, en vous basant sur des faits scientifiques, vos connaissances et votre raisonnement. Générez uniquement les lettres correspondant aux réponses correctes (par exemple : A C D). Chaque question peut avoir une ou plusieurs réponses correctes. Les justifications ne sont pas permises.
Question: *{{question}}*
Choix:
{% for letter, option in options.items() %}
{{letter}}: {{option}}
{% endfor %}
Réponse :

Figure 7: Instruction template for zero-shot evaluation used for FrMedMCQA evaluation.

For Open-Ended Question (OEQ) evaluations, we used a standard prompt (Figure 8) for all OEQ datasets except FrClinicalQA. This dataset consists of interconnected questions about clinical cases, where the prompt (Figure 9) included the clinical case, any previous questions and the current question to be answered. This structure maintained the contextual relationship between questions within each clinical case.

Instruction Template

Veuillez lire l'instruction médicale ci-dessous et fournir une réponse adaptée à la situation décrite. Votre tâche est de répondre correctement en vous basant sur des faits scientifiques et vos connaissances. Répondez uniquement à la question posée de manière brève et concise. Faites des phrases courtes contenant la réponse, évitez les informations non essentielles et concentrez-vous sur les éléments cruciaux pour une réponse efficace et pertinente.
Instruction: *{{question}}*
Réponse :

Figure 8: Instruction template for zero-shot evaluation used for OEQ evaluation.

Instruction Template

Vous allez lire un cas clinique suivi de plusieurs questions liées. Votre tâche est de répondre correctement à la dernière question en utilisant uniquement le contexte clinique fourni et les questions précédentes. N'incluez pas d'informations non pertinentes ou de réponses aux questions précédentes. Répondez de manière brève et concise à la question posée, en vous basant sur le cas clinique et les questions précédentes comme contexte.
Cas Clinique: *{{clinical_case}}*
{{previous_questions}}
Répondez uniquement à la question suivante, en utilisant le cas clinique et les questions précédentes comme contexte, sans inclure de réponses précédentes.
Question: *{{question}}*
Réponse:

Figure 9: Instruction template for zero-shot evaluation used for FrClinicalQA dataset.

F Few-shot Evaluation Analysis

To investigate whether the language of in-context examples affects model performance, we conducted 3-shot evaluations on translated MCQ tasks using two configurations: French prompts with En-

Model	Strategy	Average	PubMedQA	MedQA 4 Options	MedQA 5 Options	MedMCQA	MMLU					
							Clinical Knowledge	Medical Genetics	Anatomy	Pro. Medicine	College Biology	College Medicine
French 3-shot												
Mistral-7B-v0.1	Base Model	30.19	62.80	38.47	29.82	11.27	25.41	34.00	21.23	34.31	22.22	22.35
Mistral-7B-Nachos	CPT	35.63	**65.87**	38.39	28.91	26.31	27.30	41.67	29.88	32.35	36.57	29.09
Mistral-7B-Nachos-instruct	CPT+SFT	44.75	61.93	39.46	32.42	37.30	40.88	**58.67**	39.51	**55.51**	40.74	41.04
MedMistral-7B-chat	SFT	**46.35**	61.67	**40.32**	**32.68**	**37.37**	**47.30**	54.00	**45.68**	49.14	**45.60**	**49.71**
Mistral-7B-Instruct-v0.1	Base Model	43.40	**68.07**	31.29	27.00	34.35	49.94	50.00	42.22	**45.96**	39.35	45.86
Mistral-7B-Instruct-Nachos	CPT	**45.28**	43.73	38.15	32.18	**37.64**	**53.21**	**63.00**	**42.96**	44.12	**51.39**	**46.44**
Mistral-7B-Instruct-Nachos-instruct	CPT+SFT	42.24	60.13	**39.57**	**32.70**	35.40	40.50	53.67	39.75	40.56	40.28	39.88
BioMistral-7B	Base Model	**46.02**	**70.87**	34.85	29.77	**36.62**	**53.33**	**54.33**	**42.96**	44.12	42.82	**50.48**
BioMistral-Nachos-7B	CPT	40.37	29.73	**36.79**	**31.61**	35.80	48.81	53.33	34.32	**44.85**	**43.75**	44.70
BioMistral-Nachos-7B-instruct	CPT+SFT	38.82	38.80	35.59	31.32	35.96	36.98	47.67	37.53	42.28	40.28	41.81
BioMistral-7B-chat	SFT	38.67	49.47	35.30	28.70	33.44	46.04	42.33	42.72	38.85	32.64	37.19
English 3-shot												
Mistral-7B-v0.1	Base Model	32.36	**68.60**	38.49	31.03	11.40	28.30	34.67	21.48	39.58	26.16	23.89
Mistral-7B-Nachos	CPT	44.09	67.07	40.30	33.07	34.33	40.38	57.67	39.01	47.18	42.82	39.11
Mistral-7B-Nachos-instruct	CPT+SFT	**47.40**	66.13	**41.16**	**34.41**	**37.84**	42.64	**58.33**	42.96	**56.13**	**46.06**	**48.36**
MedMistral-7B-chat	SFT	46.47	62.07	39.85	32.47	37.61	**48.81**	54.00	**48.64**	48.90	43.98	**48.36**
Mistral-7B-Instruct-v0.1	Base Model	43.93	**68.87**	33.41	28.80	36.19	49.56	46.67	41.98	45.59	40.97	47.21
Mistral-7B-Instruct-Nachos	CPT	**48.29**	63.27	**41.21**	34.22	**39.39**	**52.20**	**63.33**	41.73	**47.30**	**51.85**	**48.36**
Mistral-7B-Instruct-Nachos-instruct	CPT+SFT	44.98	60.07	41.06	**34.93**	37.29	46.92	55.00	**42.22**	45.83	43.29	43.16
BioMistral-7B	Base Model	**46.76**	**69.93**	**37.89**	31.42	**38.79**	**53.33**	**50.67**	42.22	44.98	46.30	**52.02**
BioMistral-Nachos-7B	CPT	42.53	38.53	37.60	**33.91**	37.75	50.44	50.33	35.56	**47.92**	**48.15**	45.09
BioMistral-Nachos-7B-instruct	CPT+SFT	37.87	32.47	37.29	31.40	37.00	37.61	45.33	35.31	41.05	42.13	39.11
BioMistral-7B-chat	SFT	39.14	48.13	36.50	30.16	33.41	45.41	42.67	**45.19**	40.07	30.32	39.50

Table 8: Performance comparison of 3-shot in-context learning evaluations. The prompt is consistently in French, while the example shots are presented either in French or English. Scores are reported using exact-match accuracy. The best-performing model within each group is highlighted in **bold**, and the overall best-performing model is underlined.

glish examples (English 3-shot) and fully French prompts and examples (French 3-shot). Table 8 shows the results of these experiments, accompanied by statistical significance analysis (Table 9).

The results reveal several interesting patterns in language configuration impact. Models adapted through CPT, particularly Mistral-7B-Instruct-Nachos and Mistral-7B-Nachos, show slightly better performance when provided with English examples compared to French ones. For instance, Mistral-7B-Nachos achieves 44.09% accuracy with English examples versus 35.63% with French examples (Table 8). This pattern is particularly interesting given that these models were trained only on French medical data (NACHOS), suggesting that the base model's general English capabilities might contribute to better utilization of English examples even in medical contexts.

Group-specific performance patterns remain consistent with our main findings. Group 1 (Mistral-based) maintains its superior performance in few-shot settings, while models from Group 3 (BioMistral-based) continue to show limited improvements over their base model. The relative ranking of adaptation strategies remains stable across both few-shot configurations and aligns with zero-shot results.

These findings provide additional context to our main conclusions about adaptation strategies while revealing an advantage of English examples in few-shot scenarios, despite the models' French medical training.

G Statistical Significance Assessment Results

The statistical significance analysis results are shown across different evaluation settings: Tables 11, 12 and 13 present the win/tie/loss rates for translated MCQs, native MCQs, and OEQs, respectively. Each rate indicates the proportion of datasets where a model shows statistically significant improvement (win), no significant difference (tie), or significant degradation (loss) compared to its base model.

H Computational Resources and Environmental Impact

Table 10 details the computational resources required for each adaptation strategy and their environmental impact. We report training time, GPU requirements, carbon emissions (kgCO2e), and associated costs in USD. Carbon emissions were calculated based on the energy consumption of differ-

			Average			PubMedQA			MedQA 4 Options			MedQA 5 Options			MedMCQA			MMLU																
																		Clinical Knowledge			Medical Genetics			Anatomy			Pro. Medicine			College Biology			College Medicine		
Model	Base Model	Strategy	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss
French 3-shot																																		
Mistral-7B-Nachos	**Mistral-7B-v0.1**	CPT	**0.3**	0.7	0	0	1	0	0	1	0	0	1	0	**1**	0	0	0	1	0	0	1	0	**1**	0	0	0	1	0	**1**	0	0	0	1	0
Mistral-7B-Nachos-instruct	**Mistral-7B-v0.1**	CPT+SFT	**0.7**	0.3	0	0	1	0	0	1	0	0	1	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0
Mistral-7B-Nachos-instruct	**Mistral-7B-Nachos**	CPT+SFT	**0.5**	0.5	0	0	1	0	0	1	0	0	1	0	**1**	0	0	**1**	0	0	**1**	0	0	0	1	0	**1**	0	0	0	1	0	**1**	0	0
Med-Mistral-7B-chat	**Mistral-7B-v0.1**	SFT	**0.7**	0.3	0	0	1	0	0	1	0	0	1	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0
Mistral-7B-Instruct-Nachos	**Mistral-7B-Instruct-v0.1**	CPT	**0.4**	0.5	0.1	0	0	**1**	**1**	0	0	**1**	0	0	**1**	0	0	0	1	0	0	1	0	0	1	0	0	1	0	**1**	0	0	0	1	0
Mistral-7B-Instruct-Nachos-instruct	**Mistral-7B-Instruct-Nachos**	CPT+SFT	0.1	0.6	**0.3**	**1**	0	0	0	1	0	0	1	0	0	0	**1**	0	0	**1**	0	1	0	0	1	0	0	1	0	0	0	**1**	0	1	0
Mistral-7B-Instruct-Nachos-instruct	**Mistral-7B-Instruct-v0.1**	CPT+SFT	0.2	0.6	0.2	0	0	**1**	**1**	0	0	**1**	0	0	0	1	0	0	0	**1**	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0
BioMistral-Nachos-7B	**BioMistral-7B**	CPT	0	0.9	**0.1**	0	0	**1**	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0
BioMistral-Nachos-7B-instruct	**BioMistral-7B**	CPT+SFT	0	0.8	**0.2**	0	0	**1**	0	1	0	0	1	0	0	1	0	0	0	**1**	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0
BioMistral-Nachos-7B-instruct	**BioMistral-Nachos-7B**	CPT+SFT	0.1	0.8	0.1	**1**	0	0	0	1	0	0	1	0	0	1	0	0	0	**1**	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0
BioMistral-7B-chat	**BioMistral-7B**	SFT	0	0.7	**0.3**	0	0	**1**	0	1	0	0	1	0	0	0	**1**	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	0	**1**
English 3-shot																																		
Mistral-7B-Nachos	**Mistral-7B-v0.1**	CPT	**0.7**	0.3	0	0	1	0	0	1	0	0	1	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0
Mistral-7B-Nachos-instruct	**Mistral-7B-v0.1**	CPT+SFT	**0.7**	0.3	0	0	1	0	0	1	0	0	1	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0
Mistral-7B-Nachos-instruct	**Mistral-7B-Nachos**	CPT+SFT	**0.1**	0.9	0	0	1	0	0	1	0	0	1	0	**1**	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0
Med-Mistral-7B-chat	**Mistral-7B-v0.1**	SFT	**0.6**	0.3	0.1	0	0	**1**	0	1	0	0	1	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	0	1	0	**1**	0	0	**1**	0	0
Mistral-7B-Instruct-Nachos	**Mistral-7B-Instruct-v0.1**	CPT	**0.4**	0.5	0.1	0	0	**1**	**1**	0	0	**1**	0	0	**1**	0	0	0	1	0	**1**	0	0	0	1	0	0	1	0	0	1	0	0	1	0
Mistral-7B-Instruct-Nachos-instruct	**Mistral-7B-Instruct-Nachos**	CPT+SFT	0	0.9	**0.1**	0	1	0	0	1	0	0	1	0	0	0	**1**	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0
Mistral-7B-Instruct-Nachos-instruct	**Mistral-7B-Instruct-v0.1**	CPT+SFT	**0.2**	0.7	0.1	0	0	**1**	**1**	0	0	**1**	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0
BioMistral-Nachos-7B	**BioMistral-7B**	CPT	0	0.9	**0.1**	0	0	**1**	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0
BioMistral-Nachos-7B-instruct	**BioMistral-7B**	CPT+SFT	0	0.7	**0.3**	0	0	**1**	0	1	0	0	1	0	0	1	0	0	0	**1**	0	1	0	0	1	0	0	1	0	0	1	0	0	0	**1**
BioMistral-Nachos-7B-instruct	**BioMistral-Nachos-7B**	CPT+SFT	0	0.9	**0.1**	0	1	0	0	1	0	0	1	0	0	1	0	0	0	**1**	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0
BioMistral-7B-chat	**BioMistral-7B**	SFT	0	0.6	**0.4**	0	0	**1**	0	1	0	0	1	0	0	0	**1**	0	1	0	0	1	0	0	1	0	0	1	0	0	0	**1**	0	0	**1**

Table 9: The 3-shot win/tie/loss rates for all medical comparisons on translated MCQ benchmarks. For each medical model, we **boldface** the win rate if it wins more than it loses to its base model, and vice versa.

ent GPU types and training durations.

Model	Strategy	Dataset size (GB)	Type of GPU	Memory per GPU (GB)	Number of GPUs	Training time (hours)	Emissions (KgCO2e)	Cost (USD)
Mistral-7B-Nachos	CPT	7.4	NVIDIA H100	80	32	12	9.86	643.64
Mistral-7B-Nachos-instruct	CPT+SFT	7.4+ 0.036	NVIDIA H100/A100	80	32 + 1	12 + 75	9.86 + 1.92	643.64 + 63.22
MedMistral-7B-chat	SFT	0.036	NVIDIA A40	48	2	53	2.62	44.57
Mistral-7B-Instruct-Nachos	CPT	7.4	NVIDIA A100	80	32	40	32.89	1072.74
Mistral-7B-Instruct-Nachos-instruct	CPT+SFT	7.4+ 0.036	NVIDIA A100/H100	80	32 + 1	40 + 42	32.89 + 1.07	1072.74 + 70.48
BioMistral-Nachos-7B	CPT	7.4	NVIDIA H100	80	32	11	9.04	589.75
BioMistral-Nachos-7B-instruct	CPT+SFT	7.4+ 0.036	NVIDIA H100	80	32 + 1	11 + 42	9.04 + 1.07	589.75 + 70.48
BioMistral-7B-chat	SFT	0.036	NVIDIA L40	48	2	52	2.57	43.53

Table 10: Resource requirements and environmental impact for different adaptation strategies. Training times and costs are reported per adaptation strategy and base model. Carbon emissions are calculated based on GPU energy consumption during training.

			Average			PubMedQA			MedQA 4 Options			MedQA 5 Options			MedMCQA			MMLU														
																		Clinical Knowledge			Medical Genetics			Anatomy			Pro. Medicine			College Biology			College Medicine		
Model	Base Model	Strategy	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss
Mistral-7B-Nachos	**Mistral-7B-v0.1**	CPT	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0
Mistral-7B-Nachos-instruct	**Mistral-7B-v0.1**	CPT+SFT	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0
Mistral-7B-Nachos-instruct	**Mistral-7B-Nachos**	CPT+SFT	**0.5**	0.5	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	0	1	0	0	1	0	0	1	0	**1**	0	0	0	1	0	0	1	0
Med-Mistral-7B-chat	**Mistral-7B-v0.1**	SFT	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0
Mistral-7B-Instruct-Nachos	**Mistral-7B-Instruct-v0.1**	CPT	**0.3**	0.6	0.1	0	0	**1**	**1**	0	0	**1**	0	0	**1**	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0
Mistral-7B-Instruct-Nachos-instruct	**Mistral-7B-Instruct-Nachos**	CPT+SFT	**0.1**	0.9	0	**1**	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0
Mistral-7B-Instruct-Nachos-instruct	**Mistral-7B-Instruct-v0.1**	CPT+SFT	**0.3**	0.7	0	0	1	0	**1**	0	0	**1**	0	0	**1**	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0
BioMistral-Nachos-7B	**BioMistral-7B**	CPT	0	0.9	**0.1**	0	0	**1**	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0
BioMistral-Nachos-7B-instruct	**BioMistral-7B**	CPT+SFT	0	0.8	**0.2**	0	0	**1**	0	1	0	0	1	0	0	1	0	0	0	**1**	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0
BioMistral-Nachos-7B-instruct	**BioMistral-Nachos-7B**	CPT+SFT	**0.2**	0.8	0	**1**	0	0	0	1	0	0	1	0	**1**	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0
BioMistral-7B-chat	**BioMistral-7B**	SFT	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0

Table 11: The 0-shot win/tie/loss rates for all medical comparisons on translated MCQ benchmarks. For each medical model, we **boldface** the win rate if it wins more than it loses to its base model, and vice versa.

			Average			FrBMedQA			FrMedMCQA		
Model	Base Model	Strategy	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss
Mistral-7B-Nachos	**Mistral-7B-v0.1**	CPT	**1**	0	0	**1**	0	0	**1**	0	0
Mistral-7B-Nachos-instruct	**Mistral-7B-v0.1**	CPT+SFT	**1**	0	0	**1**	0	0	**1**	0	0
Mistral-7B-Nachos-instruct	**Mistral-7B-Nachos**	CPT+SFT	0	1	0	0	1	0	0	1	0
Med-Mistral-7B-chat	**Mistral-7B-v0.1**	SFT	**1**	0	0	**1**	0	0	**1**	0	0
Mistral-7B-Instruct-Nachos	**Mistral-7B-Instruct-v0.1**	CPT	**0.5**	0.5	0	**1**	0	0	0	1	0
Mistral-7B-Instruct-Nachos-instruct	**Mistral-7B-Instruct-Nachos**	CPT+SFT	0	1	0	0	1	0	0	1	0
Mistral-7B-Instruct-Nachos-instruct	**Mistral-7B-Instruct-v0.1**	CPT+SFT	**0.5**	0.5	0	**1**	0	0	0	1	0
BioMistral-Nachos-7B	**BioMistral-7B**	CPT	0	1	0	0	1	0	0	1	0
BioMistral-Nachos-7B-instruct	**BioMistral-7B**	CPT+SFT	0	1	0	0	1	0	0	1	0
BioMistral-Nachos-7B-instruct	**BioMistral-Nachos-7B**	CPT+SFT	0	1	0	0	1	0	0	1	0
BioMistral-7B-chat	**BioMistral-7B**	SFT	0	1	0	0	1	0	0	1	0

Table 12: The 0-shot win/tie/loss rates for all medical comparisons on 2 native french MCQ datasets. For each medical model, we **boldface** the win rate if it wins more than it loses to its base model, and vice versa.

			Average			FrClinicalQA			FrMedQA			K-QA		
Model	Base Model	Strategy	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss
Mistral-7B-Nachos	**Mistral-7B-v0.1**	CPT	0.33	0.33	0.33	**1**	0	0	0	1	0	0	0	**1**
Mistral-7B-Nachos-instruct	**Mistral-7B-v0.1**	CPT+SFT	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0
Mistral-7B-Nachos-instruct	**Mistral-7B-Nachos**	CPT+SFT	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0
Med-Mistral-7B-chat	**Mistral-7B-v0.1**	SFT	0.33	0.33	0.33	0	0	**1**	0	1	0	**1**	0	0
Mistral-7B-Instruct-Nachos	**Mistral-7B-Instruct-v0.1**	CPT	0	0.33	**0.67**	0	0	**1**	0	**1**	0	0	0	**1**
Mistral-7B-Instruct-Nachos-instruct	**Mistral-7B-Instruct-Nachos**	CPT+SFT	**0.67**	0.33	0	**1**	0	0	0	**1**	0	1	0	0
Mistral-7B-Instruct-Nachos-instruct	**Mistral-7B-Instruct-v0.1**	CPT+SFT	0	0.67	**0.33**	0	1	0	0	0	**1**	0	1	0
BioMistral-Nachos-7B	**BioMistral-7B**	CPT	0.33	0.33	0.33	**1**	0	0	0	1	0	0	0	**1**
BioMistral-Nachos-7B-instruct	**BioMistral-7B**	CPT+SFT	**1**	0	0	**1**	0	0	**1**	0	0	**1**	0	0
BioMistral-Nachos-7B-instruct	**BioMistral-Nachos-7B**	CPT+SFT	**0.67**	0.33	0	**1**	0	0	0	1	0	**1**	0	0
BioMistral-7B-chat	**BioMistral-7B**	SFT	**0.67**	0.33	0	**1**	0	0	0	1	0	**1**	0	0

Table 13: The 0-shot win/tie/loss rates for all medical comparisons on 3 OEQ datasets. For each medical model, we **boldface** the win rate if it wins more than it loses to its base model, and vice versa.